Matrix and Tensor Factorization Methods for Toxicogenomic Modeling and Prediction

Khan, Suleiman A.

Springer International Publishing AG
2019-05

Khan , S A , Aittokallio , T , Scherer , A , Grafström , R C & Kohonen , P 2019 , Matrix and Tensor Factorization Methods for Toxicogenomic Modeling and Prediction . in H Hong (ed.) , Advances in Computational Toxicology : Methodologies and Applications in Regulatory Science . Challenges and Advances in Computational Chemistry and Physics , vol. 30 , Springer International Publishing AG , Cham , pp. 57-74 . https://doi.org/10.1007/978-3-030-16443-0_4

http://hdl.handle.net/10138/330149
https://doi.org/10.1007/978-3-030-16443-0_4

acceptedVersion

Chapter 4

Matrix and Tensor Factorization Methods for Toxicogenomic Modeling and Prediction

Suleiman A. Khan [1], Tero Aittokallio [1,2], Andreas Scherer [1], Roland Grafström [3,4,5], Pekka Kohonen [3,4,5]

[1] Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

[2] Department of Mathematics and Statistics, University of Turku, Turku, Finland

[3] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

[4] Predictomics AB, Stockholm, Sweden

[5] Misvik Biology Oy, Turku, Finland

Corresponding author:

Suleiman A. Khan, Institute for Molecular Medicine Finland, University of Helsinki.

Email: khan.suleiman@gmail.com; suleiman.khan@helsinki.fi

Running Head: Advanced Machine Learning methods for Toxicogenomics

Abstract: Prediction of unexpected, toxic effects of compounds is a key challenge in computational toxicology. Machine learning based toxicogenomic modelling opens up a systematic means for genomics-driven prediction of toxicity, which has the potential also to unravel novel mechanistic processes that can help to identify underlying links between the molecular makeup of the cells and their toxicological outcomes. This chapter describes the recent big-data and machine learning-driven computational methods and tools that enable one to address these key challenges in computational toxicogenomics, with a particular focus on matrix and tensor factorization approaches. Here we describe these approaches by using exemplar application of a dataset comprising over 2.5×10^8 data points and 1300 compounds, with the aim of explaining dose-dependent cytotoxic effects by identifying hidden factors/patterns captured in transcriptomics data with links to structural fingerprints of the compounds. Together transcriptomics and structural data are able to predict pathological states in liver and drug toxicity.

Key words: Machine Learning, Group Factor Analysis, Tensor Factorization, Bayesian Modelling, Drug Sensitivity, Connectivity Map, NCI-60, Gene expression, Biomarkers;

1. Introduction

Cellular responses to drugs and other chemical compounds are increasingly being measured at multiple levels of detail and resolution. For instance, *ex-vivo* toxicity measurements summarize the phenotypic responses in human primary cells [1][2], while profiling of genome-wide transcriptomic responses opens up a systems-level view to the compounds' mode-of-action (MoA) mechanisms. *The study of relationships between genome-wide genomic or molecular responses of the cells to exposure to substances and the corresponding toxicological outcomes is referred to as toxicogenomics.* Understanding these complex relationships can not only identify the molecular mechanisms behind toxicity but also suggest ways to avoid toxic effects in medical or other applications [3]. Toxicogenomics may be especially pertinent for analyzing data from cellular assays, and for reducing and eventually replacing the use of animal experiments for toxicity testing during drug development, also referred to as 3R approaches [3][4][6]. The reductions in the costs of genomics and transcriptomic assays are enabling factors towards 3R as well [6][8].

In vitro toxicological outcomes are often based on large-scale compound response profiles, which summarize the responses in a particular cell context. For instance, NCI-60 developmental therapeutics program uses several metrics to quantify dose-responses to a library of thousands of compounds across a panel of 59 human tumor cell lines; such summary metrics include: GI50 (50% Growth Inhibition), TGI (Total Growth Inhibition) and LC50 (50% Lethal Concentration) (https://dtp.cancer.gov/discovery_development/nci-60/). In such a high-throughput setting, the computational task is to search for patterns of toxicity outcomes in correlation with genomic and molecular profiles of the same panel of cell lines. However, cytotoxicity is not a biologically uniform response. Cells use multiple mechanisms that depend

on the chemical or drug and the dose at which it is applied to respond to and counter the effects of stressors. Transcriptomic profiling and subsequent analyses using component modelling approaches discussed herein can segment these responses into biologically intelligible and explainable sub-responses, while also providing predictive models.

Therefore, advances in machine learning methodology allow study of toxicogenomic relationships in a more systematic fashion and reveal valuable drug-gene associations. For instance, community efforts have shown great promise to improve in-silico predictions of drug sensitivity [9]. In another effort, [10] carried out a personalized quantitative structure activity relationship QSAR analysis by integrating gene expression, drug structures and drug response profiles using a non-linear machine learning approach. Their study demonstrated the possibility to predict the drug sensitivity outcome for untested drugs even in new cell-types. Drug-pathway associations can be identified using advanced machine learning methodologies that model the complex molecular interactions [11]. Recently, integrative multi-task sparse regression methods have been used to systematically identify biomarker-combinations for predicting drug outcomes [12]. Increasing evidence from recent studies thus poses the hypothesis that common patterns in the activity profiles of genes and sensitivity/toxicity profiles of drugs can identify cellular response mechanisms and could be used even in predicting the tissue type or cell context-specific toxicity outcome of drug treatment.

This Chapter is organized as follows: Section 2 introduces representative classes of recent machine learning methods, with a specific emphasis on the matrix and tensor factorization methods. Section 3 demonstrates the application of these methods to identification of toxicogenomic relationships in example case studies, followed by a discussion in Section 4.

Section 5 concludes the Chapter with current limitations and future directions in these developments.

2. Machine learning methods

Machine learning algorithms search for patterns in data to extract useful information [13][14]. These algorithms learn a representation a.k.a. the model from existing data samples and then utilize the model in different tasks. When applied to experimental data, the model formulation and learning processes take into account various different forms of inherent noise and corruptions in the measurements to learn a cleaner representation of the data. In toxicogenomics, similar to many other real-life applications, data is expected to be noisy, high dimensional, contain missing values, and may also include correlated variables, which all make direct analysis complicated [15]. In such cases, the key feature of machine learning is to identify a low-dimensional, hidden representation that captures and summarizes the relevant information for the toxicogenomic modelling task. These summaries can then be used to understand the compound's MoA and/or predict the cellular outcomes of the drugs in different cell contexts.

2.1 Matrix Factorization

In machine learning, matrix factorization (MF) is a well-established approach to summarize a dataset through unobserved features that explain why some parts of the data are similar. MF has applications in broad range of scientific domains, and it is widely used in several applications, including prediction of missing values, dimensionality reduction, as well as data visualization [16][17]. This wide applicability comes from the assumption that MF can be seen as a means of describing the underlying processes which generated the data. Specifically, MF

assumes that measurements have been produced by a combination of a number of latent processes and aims to identify the factors (a.k.a. components) that describe these processes. Fig. 4.1 shows a visual illustration of matrix factorization, where a matrix **X** is factorized into distinct low-dimensional components. This component decomposition is valuable for many applications, as the different components can be related to separate mechanisms that may have contributed to the data. Several matrix factorization methods have been proposed for various applications, including Factor Analysis (FA), Principal Component Analysis (PCA) and Latent Dirichlet Allocation (LDA, see 2.2) [18][20]. While FA and PCA are designed for continuous data sets, LDA is formulated for discrete datasets.

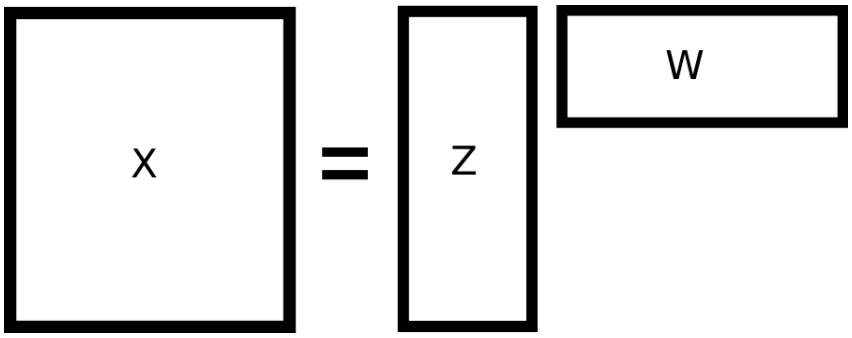

Fig. 4.1: Visual representation of matrix factorization. The data matrix X is factorized into low-dimensional matrices Z and W that capture the key statistical patterns in the data.

2.2. Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA, [20][21]) is a probabilistic formulation of factorization for discrete data sets. Formally, it is a three-level hierarchical Bayesian model that models the probabilities of each input feature to appear in each component. The Dirichlet distribution is a multivariate probability distribution used in LDA to mitigate overfitting and to help LDA to achieve its generalizability beyond the training data. LDA has demonstrated wide applicability in natural language processing, as text data sets can directly be encoded as discrete variables [22][24] as well as in genomic data sets [25][27].

2.3. Group Factor Analysis

Group Factor Analysis (GFA, [28][30]) is a recent machine learning method designed to capture relationships between multiple datasets. GFA models the relationships as statistical dependencies by reducing multiple data sets (also known as views) to learn a joint low-dimensional representation. The joint representation of the datasets is characterized by components that may be active in one or several of the data views as shown in Fig. 4.2. An active component captures underlying relationships between the views in which it is active. For example, the active component of all views captures a common dependency structure between all views, while a component active only in a single view identifies the variance and features unique to that particular view only. GFA learns the components and their activity patterns in a truly data-driven fashion, making it possible to comprehensively capture the interdependencies between all the data views. An easy to use implementation of GFA has been made freely available as an R-package [31].

Formally, for a given collection of M datasets $\mathbf{X}^{(m)} \in \mathcal{R}^{N \times D_m}$ where m = 1… M, having N paired samples and D_m dimensions, GFA learns a joint low-dimensional factorization of the M matrices. The model is formulated as a product of the Gaussian latent variable matrix $\mathbf{Z} \in \mathcal{R}^{N \times K}$ (containing the K components) and view-specific projection weights $\mathbf{W}^{(m)} \in \mathcal{R}^{D_m \times K}$:

$$\mathbf{x}_n^{(m)} \sim \mathcal{N}\left(\mathbf{W}^{(m)} \mathrm{z}_n , \sum^{(m)} \right),$$

$$\mathbf{z}_n \sim \mathcal{N}(0, \mathbf{I})$$

$$\mathrm{w}_{d,k}^{(m)} \sim h_{m,k}\, \mathcal{N}\left(0, (\alpha_{d,k}^{(m)})^{-1}\right) + (1 - h_{m,k})\delta_0$$

$$h_{m,k} \sim Bernoulli(\pi_k)$$

$$\pi_k \sim Beta(a^{\pi}, b^{\pi})$$

$$\alpha_{d,k}^{(m)} \sim Gamma(a^{\alpha}, b^{\alpha})$$

Here, $\sum^{(m)}$ is a diagonal noise covariance matrix. The latent variable $\mathbf{z}_n$ are common between all the views and capture the response patterns. The projection matrices $\mathbf{w}_{:,k}^{(m)}$ are specific to each view and translate the dependency patterns across views.

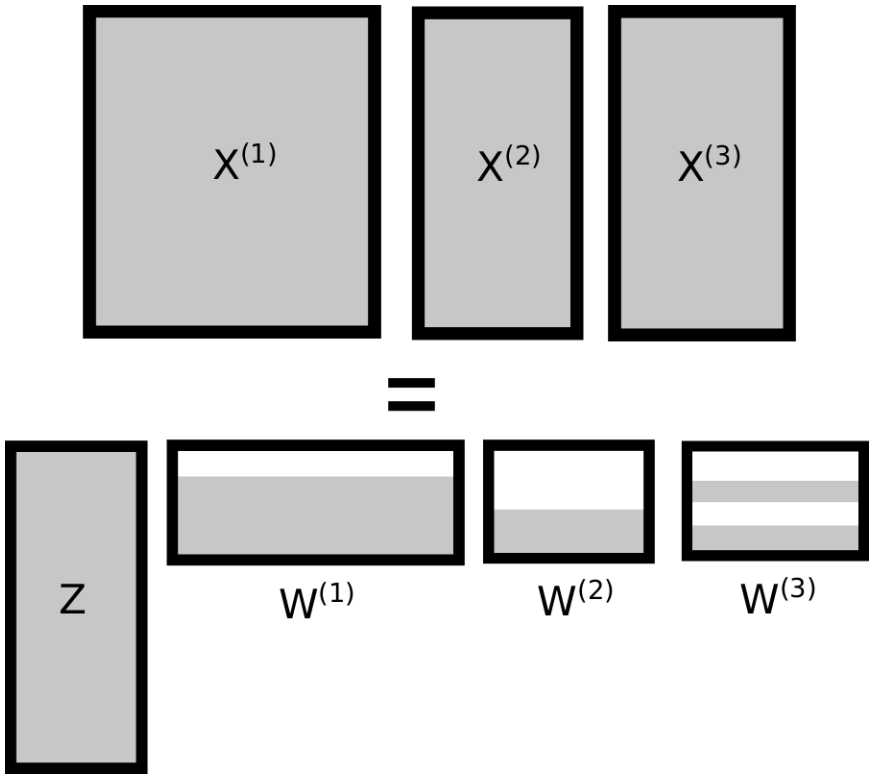

Fig. 4.2: Visual representation of Group Factor Analysis. GFA factorizes a set of data matrices **X$^{(1)}$, X$^{(2)}$.. X$^{(m)}$**, into their joint low dimensional factors **Z**. The factors can be active in one or more data matrices through the projection matrices **W$^{(1)}$, W$^{(2)}$.. W$^{(m)}$**. The **W**'s are learned to hold a group-wise sparse structure that models the dependency patterns across the data matrices. The sparsity is illustrated by white color which represents zero weights, while shaded color represents non-zero values in the figure.

GFA achieves the joint factorization by assuming that the projections $\mathbf{w}_{:,k}^{(m)}$ are group-wise sparse. The group sparse projections $\mathbf{w}_{:,k}^{(m)}$ capture both group-specific variations (activity displayed only in one view) as well as dependencies between multiple groups (activity in more

than one view). The sparsity is implemented in two layers through a group-wise group spike and slab prior formulation using Beta-Bernoulli distribution [30] and an element-wise normal-Gamma Automatic Relevance Determination (ARD) [32]. As a result, the project matrices $\mathbf{W}^{(m)}$ are both group and feature-wise sparse, which is compatible with the biological assumptions of targeted action mechanisms making the results easier to interpret.

2.4. Tensor Factorization

A tensor is a multidimensional array $\mathcal{X} \in \mathcal{R}^{I_1 \times I_2 \times \ldots \times I_j}$ and a generalization of matrices and vectors to higher order spaces. Tensors are therefore useful for representing data that has more than two dimensions. Such representation allows investigation of relationships that span multidimensional constructs. Mathematically, a tensor is also commonly defined as an element of space induced by the tensor product of vector spaces.

In order to capture the highly-structured patterns of a multidimensional data set, tensor methods employ constrained formulations that help to avoid the overfitting problem [33]. A key characteristic of these tensor formulations is that they have fewer parameters than their matrix counterparts. Analogous to matrix factorizations presented in Section 2.2, there exists several tensor factorization methods that can be used to discover underlying dependencies in the data [33]. CANDECOMP/PARAFAC (CP; [34][35]) and Tucker family [36] are the most widely used tensor decomposition methods. The interested reader is referred to [33] for a comprehensive review of various tensor factorization methods.

Tensor factorizations have obtained significant success in a large number of domains, including chemometrics, psychometrics, bioinformatics, and have shown immense promise for advanced applications in toxicology and toxicogenomics. For example, tensor factorization has

been used to explore stimuli-variant gene expression patterns [37], as well as in integrating phenotypic responses from multiple studies [38][39], modeling dependencies between metabolic and gene expression networks [40], as well as in joint QSAR and toxicogenomic analysis [41][42].

CP factorization, also known as the canonical decomposition or parallel factor analysis [34][35], is the most widely used tensor factorization method. CP is a natural extension of matrix factorization to arrays of order 3 or more as shown in Fig. 4.3. The method can be seen as carrying out simultaneous factor analysis on multiple slabs (matrices) of a tensor such that the factors of each slab differ just by a scale. CP factorization is defined in a symmetric fashion over all the modes, such that a tensor is decomposed into a sum of rank-one tensors, where each rank one tensor is the outer product of the latent vector in all modes. For a third order tensor $\mathcal{X} \in \mathcal{R}^{N \times D \times L}$, a rank-K CP is represented as:

$$\mathcal{X} = \sum_{k=1}^{K} \mathbf{z}_k \circ \mathbf{w}_k \circ \mathbf{u}_k + \epsilon$$

where **Z** and **U** and **W** are the latent variables corresponding to the three modes.

Several implementations of CP factorization have existed for quite some time now, for example the seminal implementation by [43]. Recently, CP and other factorizations have gained substantial interest amongst the machine learning community [44][45], since recent developments addressed several methodological challenges posed by multi-way data sets. More recently, an easy to use probabilistic implementation of CP was presented by [46]. The implementation automatically handles missing values in the data, hence making it applicable to a wide selection of real-world data sets. It also features automatic component selection as well as visualization and prediction routines making both exploratory and predictive analytics easier.

2.5. Multi-tensor factorization

Multi-tensor factorization (MTF, [41][42]) is a new machine learning method designed to capture relationships between a collection of tensor datasets. MTF jointly factorizes multiple tensors to learn a joint low-dimensional representation that models the statistical dependencies between the tensors. Interestingly, MTF considers matrices as tensors of order two, thus enabling joint factorization of both matrices and tensors. This characteristic makes it possible to analyze novel data sets composed of matrices as well as tensors in a single joint analysis.

MTF is designed to factorize multiple co-occurring data sets, with the objective of distinguishing the shared and specific components regardless of their matrix or tensor nature. This is achieved by modeling the entire variation of all data sets through a common Factor analysis and CP-type factorization having two keys features. First, the factorization is characterized by latent variables **Z** that are common between all the views (tensor and matrices). This allows the factorization to capture cross-dependencies regardless of the data view. Second, the loadings **W** controls which of the patterns in **Z** are active in each the views. Learning these **W** loadings makes it possible to identify the dependency patterns in a truly data-driven fashion without any prior information on dependency patterns.

Formally, for multiple paired tensors $\mathcal{X}^{(t)} \in \mathcal{R}^{N \times D_t \times L}$, where t = 1…T, we specify a joint model of matrices and tensor. An indicator variable β_t identifies the tensors (β_t=1) and matrices (β_t=2), MTF is formulated using normal distributions and conjugate priors as:

$$x^{(t)}_{n,d_t,l} \sim \mathcal{N}\left(z_{n,k} \cdot w_{d_t,k} \cdot u_{l,k} , (\tau^{(t)})^{-1} \right)$$

$$\mathbf{Z}, \mathbf{U}^{(t)} \sim \mathcal{N}(0, \mathbf{I})$$

$$w_{d,k}^{(t)} \sim h_{t,k}\, \mathcal{N}\left(0, (\alpha_{d,k}^{(t)})^{-1}\right) + (1 - h_{t,k})\delta_0$$

$$h_{t,k} \sim Bernoulli(\pi_k)$$

$$\pi_k \sim Beta(a^{\pi}, b^{\pi})$$

$$\alpha_{d,k}^{(t)} \sim Gamma(a^{\alpha}, b^{\alpha})$$

$$\tau^{(t)} \sim Gamma(a^{\tau}, b^{\tau}).$$

Here, the latent variables **Z** and **U** are common to all the tensors and capture the underlying patterns, while $\mathbf{W}^{(t)}$ translate these patterns for each tensor. The binary variables $h_{t,k}$ control the view activity through a spike and slab prior, and are automatically learned from the data. The model also enforces feature-wise sparsity through α to learn sparse features that are easier to interpret. The method is implemented using a Gibbs sampler in R programing language and made available freely (http://research.ics.aalto.fi/mi/software/MTF). The implementation learns the model parameters in a Bayesian formulation, while providing default settings for all the hyperparameters.

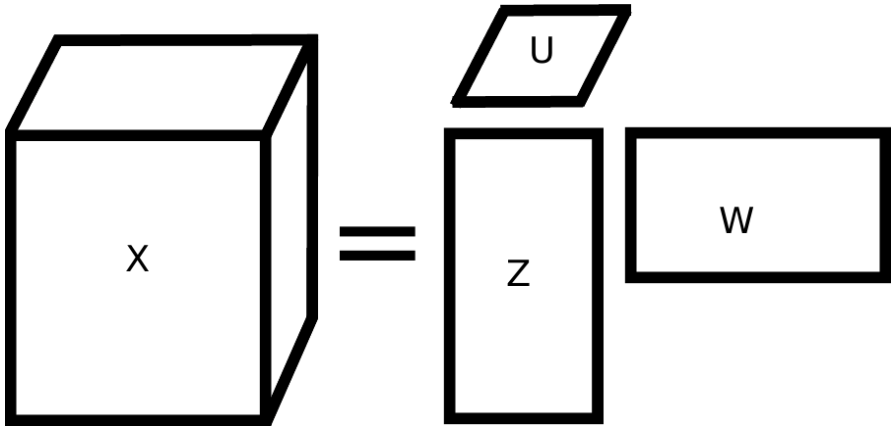

Fig. 4.3. Visual representation of CP factorization of a third order tensor. The data tensor $\mathcal{X}$ is factorized into low-dimensional matrices **Z**, **U** and **W** that capture the key statistical patterns in the data.

3. Selected case studies

3.1. Toxicogenomic Datasets

The toxicogenomic tools described in this Chapter are primarily built upon the Connectivity Map (CMap) and NCI60 data sets. CMap, introduced by the US Broad Institute, is a compendium of gene expression response profiles to 1309 small-molecules comprising mostly FDA approved drugs ([47]; https://www.broadinstitute.org/connectivity-map-cmap). The post-treatment measurements originated from three main cancer cell-lines spanning different tissues or cell-types, namely, breast (MCF7), prostate (PC3) and blood (HL60). CMap has been widely used to study interactions between small molecules, genes and diseases for various purposes including understanding the drug MoA, identifying biologically similar compounds as well as molecular mechanisms of toxicity. The treatment vs control differential gene expression (log2 readout) was obtained from the CMap dataset, such that positive expression values represent up-regulation and negative represent down-regulation as a result of treatment [4][48].

The NCI60 is a unique data repository from the US National Cancer Institute (NCI) that screened thousands of compounds over 59 cancer cell lines to provide measurements of drug responses (Shoemaker 2006; https://dtp.cancer.gov/discovery_development/nci-60). Drug response metrics include GI50 (50% Growth Inhibition), total growth inhibition (TGI), and LC50 (50% Lethal Concentration). A large number of common compounds tested in CMap cell lines were also profiled by the NCI60 program. This presents the unique opportunity to study toxic effects by integrating these two, large-scale data sets (see Section 3.3). For each CMap drug-cell pair which was also screened by NCI60, a dose-dependent toxicity score was computed such that

positive values indicated that the CMap instance is profiled at a drug-concentration higher than GI50, TGI or LC50, and therefore suggest a dose-dependent cytotoxic response.

Fig. 4.4. Toxicogenomic component activity plot. The plot shows the components that are found by the GFA model as active in the joint gene expression and toxicity data set. The y-axis shows the component number in ascending order while the x-axis shows the two data sets. The components colored black are active. The model was run for K=40 components and a total of 8 components (bottom black in both Gene expression and Toxicity) are found as shared between the two datasets. These components capture statistically patterns that are correlated across the two data sets, and are hence can be hypothesized for representing molecular mechanisms of toxicity.

3.2 Multi-view toxicogenomic using group factor analysis

To study the gene-toxicity relationships, we performed an integrated modeling of the two data sets, CMap and NCI60. The CMap data comprised detailed gene level differential expression profiles that represent the molecular response space across 11350 genes, measured after 222 drug treatments across 3 cell-lines. These data were preprocessed as previously described [48]. To focus the analysis, the 1106 highest variance genes were selected to form an expression matrix consisting of 222 drug-cell samples x 1106 genes. The toxicity values described in Section 3.1 were used to represent profiles of 222 drug-cell samples x 3 toxicity measures.

Group factor analysis (GFA) is designed to model the relationships between multiple datasets. Here, GFA was used to identify the toxicogenomic dependencies between drug-induced gene expression changes and toxicity scores. These dependencies, once identified, can elicit insights into molecular mechanisms of toxicity. GFA was run with large enough

components as specified by [28], identifying 8 shared components that capture cross-expression and toxicity relationships, as shown in Fig. 4.4, whereas a number of components found were specific to one of the datasets only. The shared components model the dependencies between the datasets while those specific to gene expression capture patterns that are not correlated with toxicity and vice-versa. The components 1 through 8 had varying numbers of genes attached to them: 518, 748, 39, 90, 27, 45, 16 and 20. The first two components included an excess of up-regulated genes (component 1: 316) and down-regulated (component 2: 706) genes, respectively.

Functional analysis of the eight components was performed with Ingenuity Pathway Analysis (IPA) which indicated that the first two components captured the largest number of biological mechanisms. The first component is highlighted here, as up-regulated genes are most informative for biomarker analysis applications (Fig. 4.5). Component 1 enriched for many organ-toxicity-related gene lists, including hepatic cholestasis and liver necrosis as well as functional pathways related to oxidative stress, the p53 pathway activation and Nf-kappa B signaling and Toll-like receptor (TLR) activation. RELA, the NF-kappa B regulator, was predicted to be most clearly effected ($p<10\text{-}16$ and Z-score 3.5). Others included the TP53 ($p<10\text{-}10$, $Z>1$), TLR-related ECSIT ($p<10\text{-}15$, $Z > 3.5$) and NR3C1 ($p<10\text{-}14$, $Z<-05$), supporting the role of NF-kappa B signaling and the Toll-like receptor activation for component 1. The GFA model identified related drugs across all three cancer types, hence suggesting a generic response of the drugs.

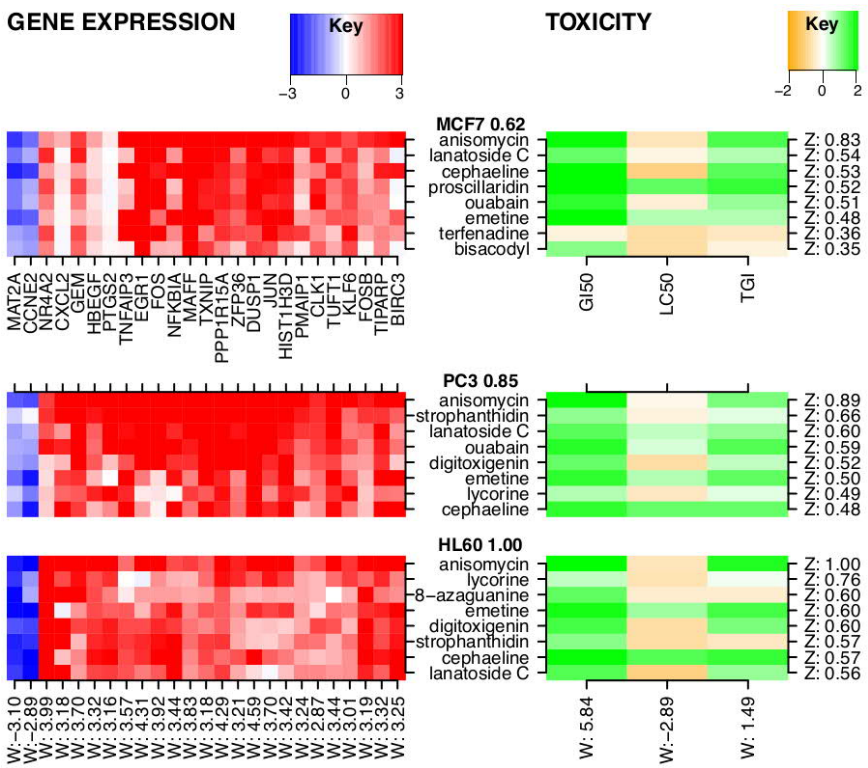

Fig 4.5. The first shared component found by GFA. The plot shows the gene expression profiles of the top genes for the top drugs of the component across the three cancer cell lines (MCF7, PC3 and HL60). Red represents upregulated expression while blue is downregulated expression. The correspondingly active toxicity profiles of the same drugs are shown on the right. Here green represents high dose-dependent toxicity values.

The second component included many cell cycle-related genes, as could be expected for a component which mainly contained down-regulated genes. Similar pathways were found activated among the 14 predictive toxicogenomic space PTGS components derived using the LDA analysis, and there is an average of almost 40% overlap between the PTGS genes and the GFA genes [4]. It is interesting to note that the first two GFA components were much larger than the other six, whereas the PTGS components had more equal numbers of genes that were significantly associated with them. Further studies would be needed to verify the utility of the GFA components for toxicity-mode-of-action studies, including biomarker discovery and Drug Induced Liver Injury (DILI) prediction.

3.3 Structural-toxicogenomic using multi-tensor factorization

Toxicogenomic applications can be extended to simultaneously include a quantitative structure activity response (QSAR) analysis, by modeling the dependencies between cellular responses of drugs and their structural descriptors. The formulation can, therefore, explore, identify and predict genomic responses linked to drugs toxicity, while simultaneously discovering their cancer specificity and correspondence to structural properties of the drugs.

Data collection for such analysis can be represented as a set of multiple tensors and matrices. In this example, we specified two tensors and one matrix. The post-treatment gene expression data from CMap was represented as the first tensor of drugs *times* cancers *times* genes dimensions. Multiple toxicity measures such as GI50, TGI and LC50 from the NCI60 formed the second tensor of drugs *times* cancers *times* toxicity measures. Finally, the structural properties of the drugs were represented as a matrix of drugs *times* descriptors.

The expression and toxicity data sets from CMap and NCI60 were processed as described in [4]. For drug structures, the modelling could make use of one or more different types of structures based on the hypothesis being tested; for example, [29] used both 3D descriptors and 2D fingerprints of the drugs for structure-response analysis. In this example, Functional Connectivity Fingerprints FCFP4 were used for representing the structural properties of the drugs. FCFP4 are advanced 2D circular topological fingerprints that have been designed for modeling of structure-activity relationships.

The multi-tensor factorization (MTF) method of Section 2.3 [42] was used to explore the structural toxicogenomic relationships. The model identified three key response components that are shared between gene expression, toxicity and structural datasets, revealing findings that are both recently established as biological insights, as well as new biological discoveries that may

have potential impact. The first component identified a response primarily driven by three Heat Shock Protein (HSP) inhibitor drugs, geldanamycin, tanespimycin, and alvespimycin, all of which are structurally analogous drugs. The drugs demonstrated an HSP response of the cells as through up-regulation of key HSP genes. This pan-cancer response across all three cancers is linked to the toxicity outcomes of the drugs. HSP90 is a molecular chaperone protein that is essential for stabilization of a variety of other proteins [50], and HSP90 inhibitors bind to the protein, resulting in its loss of function. HSP90 inhibitors have been evaluated for their therapeutic efficacy in multiple cancers [51]. This component, therefore, presents a well-known HSP90 response of cancer cells. For details on this and other components, see [42].

3.4. Predictive toxicogenomic space (PTGS)

Predictive toxicogenomics space (PTGS, [4]**)** is a recent 'big data compacting and data fusion' methodology to model various adverse and toxic outcomes on cellular and organism levels. A machine learning based data summarization approach was applied on a large transcriptomics data set. This methodology formed a predictive tool termed PTGS that used as features over 1000 genes distributed over 14 overlapping cytotoxicity-related gene space components, as described in [4]. Specifically, a LDA matrix factorization-based method was applied to the gene profiles from the Connectivity Map dataset, and the resulting summarized components were fused with cytotoxicity data from the NCI-60 cancer cell line screens to generate the PTGS. The PTGS tool was validated for predicting Drug Induced Liver Injury (DILI) and liver cytopathologal changes by calculating PTGS component-scores within three liver-related subsets of the independent TG-GATEs database [52], being the largest public toxicogenomics database. It was shown to successfully capture all the studied liver pathological changes in rats, and in conjunction with

human therapeutic drug exposure levels (C_{max}), was able to facilitate the use of cell culture-derived toxicogenomics experiments with human and rat hepatocytes to predict DILI with greater accuracy than other *in vitro* methods [4].

4. Discussion

Recent advances in machine learning methodologies have made it possible to perform integrated analysis of the gene expression response data and toxicity profiles directly. Such detailed analysis offers deeper insights by linking the activity patterns of the genes directly with the toxicity responses, and hence enriching the factor components with detailed interactions. As molecular responses of cancer cells are known to depend on a multitude of factors, including drug MoA, cell type and cellular states, simultaneous modeling of these various factors is beginning to attract attention. Specifically, in cancer, cells are known to be heterogeneous and respond selectively to targeted drugs, making it valuable to systematically model the various factors and segregate responses specific to a particular cancer-type from those which are generic. A limitation of current methods is the ability to handle missing values particularly when considering overlap between different data sets. Compared to many other "Big Data" study areas, biomedical data is less extensive and contains more missing values. The LDA method used with the PTGS had the advantage that the entire CMap datset could be used to derive the initial components, whereas the GFA method required at least some overlap between all variables, reducing the amount of gene expression data used. Tensor factorization methods are even less tolerant of missing values. Therefore unique methodological considerations and trade-offs apply to each study.

A key outcome of these joint analysis is the ability to predict the toxicity outcomes of compound treatment. The prediction of unexpected toxic effects is a challenging and important goal in toxicology. The presented first steps in computational toxicogenomic open up a systematic way for genomics-driven prediction of toxic effects. In addition, these provide novel mechanistic insights into the links between genomic measurements of cells and toxicological profiles of drugs. Gene expression response profiles of drugs present a popular systems-level view, while toxicity profiles summarize the drugs' phenotypic behavior. Large repositories of gene expression and drug sensitivity profiles such as those emerging from NCI60, CMap, CCLE, Sanger, and LINCS profile cellular responses at several levels of detail in a cell context-specific manner. With the emergence of heterogeneous and partially-paired data sets, joint factorizations are gaining popularity to identify novel dependency patterns, as well as to design powerful predictive applications [29][53]. These recent advances in machine learning, and especially the methods described in Section 2, enable systematic analysis of such large data repositories to provide novel toxicogenomic insights and predictions.

5. Conclusion and future directions

State-of-the-art machine learning methods have been presented here for modeling various toxicogenomic relationships. These advanced computational methodologies enable integration of disparate, high-dimensional data sources, including but not limited to omics, drug screening, chemical structures and drug-targets to achieve novel toxicogenomic analysis in terms of:

(i) providing means for predicting personalized toxicity outcomes,

(ii) identifying toxic modes of action, and

(iii) enabling quantitative structure activity modeling.

The here presented works suggest novel directions for future analysis. From the application perspective, matrix and tensor factorization methods can serve to stimulate integrative analysis of various toxicological and toxicogenomic datasets to suggest novel hypotheses. For example, a joint analysis of omics, toxicity and drug target data sets can help to identify disparate target-driven and toxic molecular mechanisms. Integrative analysis with drug-side effect repositories can help draw novel interactions between disease, side-effect and toxicity mechanisms. From a holistic angle, a large-scale analysis may even help us understand the different toxic states of a cell and the molecular drivers of each cellular state.

While there are limitations in the current analysis, future extensions of the analyses can advance our knowledge in various directions. First, using detailed drug-target interactions in the models could help classify the on-target and off-target effects more reliably; however, a key limitation here is to obtain large-scale standardized drug-target profiles. Very recently works in standardizing the drug-target interactions have come up on a large-scale [54] and exploring these for an integrated drug-target-toxicogenomic analysis would be an interesting future direction. Secondly, a large majority of toxicity analysis is performed on data originating from cell line panels. It would be valuable to explore if tissue-specific toxicity profiles are available for a more robust and practically applicable analysis. Third, organism-level toxicity data are limited to only a few organisms only; it is important to evaluate how comprehensive such modelling is in general and how widely the results can be applied across organisms.

In terms of future developments in the toxicology practices, studies and risk assessment strategies, we hope the presented works could stimulate the integration of advanced machine learning models. For example, the methods presented here can be used to identify the markers of toxic response towards a data and knowledge driven approach for risk assessment.

6. Reference:

[1] Grabinger T, et al. (2014) Ex vivo culture of intestinal crypt organoids as a model system for assessing cell death induction in intestinal epithelial cells and enteropathy, Cell death & disease, 5(5), e1228.

[2] Aberdam E, et al. (2017) Induced pluripotent stem cell-derived limbal epithelial cells (LiPSC) as a cellular alternative for in vitro ocular toxicity testing, PloS one, 12(6), e0179913.

[3] Hartung T, et al. (2012) Food for thought ... systems toxicology, ALTEX 29(2): 119–128.

[4] Kohonen P, et al. (2017) A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury, Nature Communications, 8, 15932.

[5] Kohonen, P. et al. (2014) Cancer biology, toxicology and alternative methods development go hand-in-hand. Basic Clin Pharmacol Toxicol. 115, 50-8.

[6] Grafström, R.C. et al. (2015) Toward the replacement of animal experiments through th bioinformatics-driven analysis of 'omics' data from human cell cultures, Altern Lab Anim. 43, 325-32.

[7] Nymark P. et al. (2018) A Data Fusion Pipeline for Generating and Enriching Adverse Outcome Pathway Descriptions. Toxicol Sci.; 162(1):264-275.

[8] Yeakley J.M. et al. (2017). A trichostatin A expression signature identified by TempO-Seq targeted whole transcriptome profiling. PLoS One; 12(5).

[9] Costello J C, et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nature biotechnology 32(12): 1202-1212.

[10] Ammad-Ud-Din M, et al. (2014). Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. Journal of chemical information and modeling 54(8): 2347-2359.

[11] Ammad-ud-din M, et al. (2016). Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. Bioinformatics 32(17): i455-i463.

[12] Ammad-ud-din M, et al. (2017) Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression Bioinformatics 33(14): i359-i368.

[13] Bishop C M, (2006) Pattern Recognition and Machine Learning. Springer, New York, USA.

[14] Gelman A, et al. (2013) Bayesian data analysis. Chapman and Hall/CRC.

[15] Bartholomew D J, et. al. (2011) Latent variable models and factor analysis: A unified approach, John Wiley & Sons, 904.

[16] Salakhutdinov R, and Mnih A, (2008) Bayesian probabilistic matrix factorization using markov chain monte carlo, in Proceedings of the 25th international conference on Machine learning, 880–887.

[17] M. E. Tipping and C. M. Bishop, (1999) Probabilistic principal component analysis, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3): 611–622.

[18] Witten D M, et al. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics, 1-8.

[19] Kossenkov A V, and Ochs M F, (2009) Matrix factorization for recovery of biological processes from microarray data, Methods in enzymology, 467: 59–77.

[20] Blei D M, et al. (2003) Latent dirichlet allocation. Journal of machine Learning research, 3: 993-1022.

[21] Ghahramani Z, (2015) Probabilistic machine learning and artificial intelligence. Nature 28, 452–459.

[22] Guo Y, et al. (2017) Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management, 59:467-483.
[23] Moro S et al. (2015) Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Systems with Applications 42(3): 1314-1324.
[24] Krestel R, et al. (2009). Latent dirichlet allocation for tag recommendation. In Proceedings of the third ACM conference on Recommender systems 61-68.
[25] Caldas J, et al. (2009) Probabilistic retrieval and visualization of biologically relevant microarray experiments, Bioinformatics 25(12):145–153.
[26] Pinoli P, et al. (2014) Latent Dirichlet allocation based on Gibbs sampling for gene function prediction. In Computational Intelligence in Bioinformatics and Computational Biology 1-8.
[27] Backenroth D et al. (2018) FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications. The American Journal of Human Genetics 102(5): 920-942.
[28] Virtanen S, et al. (2012) Bayesian group factor analysis. In Artificial Intelligence and Statistics 1269-1277.
[29] Khan S A, (2014) Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis, Bioinformatics 30(17): i497-i504.
[30] Klami A, et al. (2015). Group factor analysis. IEEE transactions on neural networks and learning systems, 26(9): 2136-2147.
[31] Leppäaho E, et al. (2017). GFA: exploratory analysis of multiple data sources with group factor analysis. The Journal of Machine Learning Research, 18(1): 1294-1298.
[32] Neal R M, Bayesian learning for neural networks, Springer-Verlag, 1996.
[33] Kolda T, and Bader B, (2009) Tensor decompositions and applications, SIAM Review 51(3): 455–500.
[34] Carroll J D, and Chang J J, (1970) Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition, Psychometrika, 35(3): 283–319.
[35] Harshman R A (1970), Foundations of the parafac procedure: models and conditions for an explanatory multimodal factor analysis, UCLA Working Papers in Phonetics, 16: 1–84.
[36] Tucker L R, (1966) Some mathematical notes on three-mode factor analysis, Psychometrika, 31(3): 279–311.
[37] Yener B, et al. (2008), Multiway modeling and analysis in stem cell systems biology, BMC systems biology, 2(1): 63.
[38] Omberg L, et al. (2007) A tensor higher-order singular value decomposition for integrative analysis of dna microarray data from different studies, Proceedings of the National Academy of Sciences 104(47): 18371–18376.
[39] Li W, et al. (2011) Integrative analysis of many weighted co-expression networks using tensor computation," PLoS computational biology, 7(6) e1001106.
[40] Brink-Jensen K, et al. (2013) Integrative analysis of metabolomics and transcriptomics data: a unified model framework to identify underlying system pathways, PloS one, 8(9) p. e72116.
[41] Khan S A, and Kaski S, (2014) Bayesian multi-view tensor factorization. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg: 656-671.
[42] Khan S A, et al. (2016) Bayesian multi-tensor factorization, Machine Learning, 105(2): 233-253.
[43] Andersson C A, and Bro R, (2000) The N-way toolbox for MATLAB, Chemometrics and Intelligent Laboratory Systems, 52(1): 1-4.
[44] Mørup M, and Hansen L K, (2009) Automatic relevance determination for multiway models, Journal of Chemometrics, 23(7) 352-363.

[45] Xiong L, (2010) Temporal collaborative filtering with bayesian probabilistic tensor factorization, in Proceedings of SIAM Data Mining, 10: 211–222.
[46] Khan S A, and Ammad-ud-din M, (2017) tensorBF: an R package for Bayesian tensor factorization, bioRxiv, 6097048 1-6.
[47] Lamb J, et al. (2006) The Connectivity Map: Using Gene-Expression Signatures to Connect Small
Molecules, Genes, and Disease. Science 313(5795): 1929-1935
[48] Khan S A, et al. (2012) Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs. BMC bioinformatics, 13(1), 112-127.
[49] Shoemaker R H, (2006) The nci60 human tumour cell line anticancer drug screen. Nature Reviews Cancer 6(10): 813–823
[50] Isaacs J S, et al. (2003), Heat shock protein 90 as a molecular target for cancer therapeutics, Cancer Cell, 3(3): 213-217.
[51] Neckers L, and Workman P, (2012) Hsp90 molecular chaperone inhibitors: are we there yet?, Clinical Cancer Research, 18(1): 64-76.
[52] Igarashi Y, et al. (2015) Open TG-GATEs: a large-scale toxicogenomics database Nucleic Acids Res. 43: D921–D927.
[53] Hore V, et al. (2016) Tensor decomposition for multiple-tissue gene expression experiments, Nature genetics, 48(9): 1094.
[54] Tang J, et al. (2018) Drug Target Commons: A Community Effort to Build a Consensus Knowledge Base for Drug-Target Interactions, Cell chemical biology, 25(2): 224-229.